Article

A Globally Convergent Iterative Method for Matrix Sign Function and Its Application for Determining the Eigenvalues of a Matrix Pencil

Munish Kansal [1], Vanita Sharma [2], Pallvi Sharma [1] and Lorentz Jäntschi [3],*

1   Department of Mathematics, Thapar Institute of Engineering and Technology, Patiala 147004, India;
    munish.kansal@thapar.edu (M.K.); psharma_phd22@thapar.edu (P.S.)
2   School of Commerce, SVKM's Narsee Monjee Institute of Management Studies, Chandigarh 160014, India;
    vanita.sharma@nmims.edu
3   Department of Physics and Chemistry, Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
*   Correspondence: lorentz.jantschi@gmail.com

Abstract: In this research article, we propose a new matrix iterative method with a convergence order of five for computing the sign of a complex matrix by examining the different patterns and symmetry of existing methods. Analysis of the convergence of the method was explored on a global scale, and attraction basins were demonstrated. In addition to this, the asymptotic stability of the scheme was explored.Then, an algorithm for determing thegeneralized eigenvalues for the case of regular matrix pencils was investigated using the matrix sign computation. We performed a series of numerical experiments using numerous matrices to confirm the usefulness and superiority of the proposed method.

Keywords: iterative method; matrix sign function; basins of attraction; generalized eigenvalue problem; matrix pencil

MSC: 65F60; 15A24

1. Introduction

Over the past several years, the efficient computing of matrix functions has emerged as a prominent area of research. Symmetry plays a significant role in the study of matrix functions, particularly in understanding their properties, behavior, and computational methods. Symmetry principles are widely used in physics and engineering to analyze and solve problems involving matrix functions. For instance, in quantum mechanics, symmetric matrices often arise in the context of Hamiltonians, and efficient computation of matrix functions is crucial for simulating quantum systems and predicting their behavior. In order to compute matrix functions, various iterative schemes have been suggested, which include methods for the evaluation of matrix sign function. The scalar sign function defined for $z \in \mathbb{C}$ such that $Re(z) \neq 0$ which is given as,

$$sign(z) = \begin{cases} 1, & Re(z) > 0, \\ -1, & Re(z) < 0. \end{cases} \tag{1}$$

is extended to form the matrix sign function. Roberts [1] introduced the Cauchy integral representation of the matrix sign function in 1980, given by

$$sign(A) = \frac{2}{\pi} A \int_0^\infty (t^2 I + A^2)^{-1} dt, \tag{2}$$

which can be applied in model reduction and in solving algebraic Riccati and Lyapunov equations. Recall the definition of matrix sign function $sign(A)$ via the Jordan canonical

form [2], in which if $A = PJP^{-1}$ is the Jordan decomposition of A such that $J = diag(J_1, J_2)$, then

$$sign(A) = P \begin{bmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{bmatrix} P^{-1}, \tag{3}$$

where $A \in \mathbb{C}^{n \times n}$ has no eigenvalue on the imaginary axis, P is a nonsingular matrix, and p and $n - p$ are sizes of Jordan blocks J_1, J_2 corresponding to eigenvalues lying in left half plane ($\mathbb{C}^-$) and right half plane ($\mathbb{C}^+$), respectively.

In the pursuit of evaluating $sign(A)$ as a solution of

$$W^2 - I = 0, \tag{4}$$

Roberts applied Newton's method (NM) resulting into the following iterative formula:

$$W_{l+1} = \frac{1}{2}(W_l + W_l^{-1}), \quad l = 0, 1, 2, \ldots, \tag{5}$$

with the initial matrix $W_0 = A$ and convergence of order two.

Another alternative is accelerated Newton's method (ANM) that can be achieved via norm scaling parameter in (5) given by,

$$\begin{cases} W_0 & = A, \\ \mu_l & = \sqrt{\frac{\|W_l^{-1}\|}{\|W_l\|}}, \\ W_{l+1} & = \frac{1}{2}(\mu_l W_l + \mu_l^{-1} W_l^{-1}). \end{cases} \tag{6}$$

In order to improve the convergence, acceleration and scaling of iteration (5), Kenney and Laub [3,4] suggested a family of matrix iterations by employing the Padé approximant to the function

$$H(x) = (1 - x)^{-1/2}, \tag{7}$$

where $x = 1 - z^2$ and taking into the consideration

$$sign(z) = \frac{z}{(1-x)^{1/2}}. \tag{8}$$

Let (r, s)-Padé approximant to $H(x)$ with $r + s \geq 1$ is denoted as $\frac{P_{r,s}(x)}{Q_{r,s}(x)}$, then

$$w_{l+1} = \frac{w_l P_{r,s}(1 - w_l^2)}{Q_{r,s}(1 - w_l^2)} = \psi_{2r+1,2s}, \tag{9}$$

converges to ± 1 if $r \geq s - 1$ and it has convergence rate equal to $r + s + 1$. So, the (r, s)-Padé approximant iteration ($PA_{r,s}$) for computing $sign(A)$ is

$$W_{l+1} = W_l[P_{r,s}(1 - W_l^2)][Q_{r,s}(1 - W_l^2)]^{-1}. \tag{10}$$

These iterations were later investigated by Gomilko et al. [5]. Other popular iterative approaches for $sign(A)$, such as the Newton-Schulz method (NSM) [6]

$$W_{l+1} = \frac{1}{2}W_l(3I - W_l^2), \tag{11}$$

and Halley's method (HM) [2]

$$W_{l+1} = [I + 3W_l^2][W_l(3I + W_l^2)]^{-1}. \tag{12}$$

are either members of the Padé family (10) or its reciprocal family.

A recent fifth order iterative method [7] (represented as Lotfi's Method (*LM*)) for $sign(A)$ can be investigated and is given by

$$W_{l+1} = [W_l(18I - 20W_l^2 - 30W_l^4)][5I + 15W_l^2 - 45W_l^4 - 7W_l^6]^{-1}. \tag{13}$$

For more recent literature on matrix sign function, one can go through references [8–10].

Motivated by constructing efficient iterative methods like the methods discussed above and to overcome on some of the drawbacks of the existing iterative methods, here we aim to propose a globally convergent iterative method for approximation of $sign(A)$. Matrix sign function has several application in scientific computing (see e.g., [11,12]). Another applications of matrix sign function include solving various matrix equations [13], solving stochastic differential equations [14] and many more.

In this article, we present a fifth order iterative scheme for evaluation of $sign(A)$. The outline for the rest of the article is defined here: In Section 2, the construction of new iterative scheme for evaluating matrix sign function has been concluded. In Section 3, basins of attraction guarantees the global convergence of developed method. Also, convergence analysis for the fifth rate of convergence is discussed along with stability analysis. Then the proposed method is extended to determine the generalized eigenvalues of a regular matrix pencil in Section 4. Computational complexity is examined in Section 5. Numerical examples are provided to evaluate the performance and to demonstrate the numerical precision of suggested method in Section 6. The final conclusion is presented in Section 7.

2. A Novel Iterative Method

In this section, we have developed an iterative method for evaluation of $sign(A)$. The primary goal is to develop the scheme by examining the factorization patterns and symmetry of existing schemes rather than using a non-linear equation solver [13]. We aim to design an iterative formula that does not conflict with the Padé family members (9) or its reciprocals.

If we assume the eigenvalues to converge to 1 or −1, and we aim to get a higher convergence order five, then the iterative formula will satisfy the equation

$$\frac{w_{l+1} - 1}{w_{l+1} + 1} = \frac{(w_l - 1)^5}{(w_l + 1)^5}. \tag{14}$$

If we solve Equation (14) for w_{l+1}, then we get

$$w_{l+1} = \frac{w_l(5 + 10w_l^2 + w_l^4)}{1 + 10w_l^2 + 5w_l^4}. \tag{15}$$

Nevertheless, the formula mentioned above is not a novel one, as it is a member of Padé family (9). So, we make some modifications in (14) so that the formula becomes better in terms of converegnce as follows:

$$\frac{w_{l+1} - 1}{w_{l+1} + 1} = \frac{(w_l - 1)^5(-4 + w_l)}{(w_l + 1)^5(-4 - w_l)}, \tag{16}$$

provided that $\left\|\frac{-4+w_l}{-4-w_l}\right\| < 1$. Solving (16) for w_{l+1}, we get

$$w_{l+1} = \frac{w_l(21 + 50w_l^2 + 9w_l^4)}{4 + 45w_l^2 + 30w_l^4 + w_l^6}, \tag{17}$$

which contains a polynomial of degree 5 in the numerator and a polynomial of degree 6 in the denominator. From (9), it is clear that Padé family has a unique member with a polynomial of degree $2r + 1$ in numerator and a polynomial of degree $2s$ in denominator for each $r, s \geq 0$. So $2r + 1 = 5$ and $2s = 6$ will result into $r = 2$ and $s = 3$, i.e., $PA_{2,3}$ and

will have order $2 + 3 + 1 = 6$ which is different from the order 5 that we are achieving. So, the above formula (17) is not a part of the family (9) or reciprocals of (9) and therefore our proposed method (PM) is as follows:

$$W_{l+1} = [W_l(21I + 50W_l^2 + 9W_l^4)][4I + 45W_l^2 + 30W_l^4 + W_l^6]^{-1}, \tag{18}$$

where W_0 is the appropriate initial guess which we discuss in Section 3.

Remark 1. *Convergence can be slow if $\|W_l\| >> 1$ for some l. In that situation, it is advisable to introduce the scaling parameter μ_l [15] given as,*

$$\mu_l = \begin{cases} \sqrt{\dfrac{\|W_l^{-1}\|}{\|W_l\|}} & \text{(norm scaling)}, \\ \sqrt{\dfrac{\rho(W_l^{-1})}{\rho(W_l)}} & \text{(spectral scaling)}, \\ \sqrt{|det(W_l)|^{-1/n}} & \text{(determinant scaling)}. \end{cases} \tag{19}$$

to speed up the convergence [16] of the scheme.

Hence, we also present the accelerated proposed method (APM) which is accelerated version of (18) given by

$$\begin{cases} W_0 & = A, \\ \mu_l & = \text{Scaling parameter computed by (19)}, \\ W_{l+1} & = [\mu_l W_l(21I + 50\mu_l^2 W_l^2 + 9\mu_l^4 W_l^4)][4I + 45\mu_l^2 Z_l^2 + 30\mu_l^4 W_l^4 + \mu_l^6 W_l^6]^{-1}, \end{cases} \tag{20}$$

where $\lim_{l\to\infty} \mu_l = 1$ and $\lim_{l\to\infty} W_l = S = sign(A)$.

3. Convergence and Stability Analysis

In this section, we discuss the convergence analysis to ensure that the sequence generated from (18) converges to $sign(A)$ for a matrix $A \in \mathbb{C}^{n \times n}$. We draw the basins of attraction to confirm the global convergence of the proposed scheme and compared with the existing schemes for solving $w^2 - 1 = 0$ [17].

We consider a region $R = [-2, 2] \times [-2, 2] \in \mathbb{C}$. The region has grid size of 256×256 and has two simple roots -1 and 1. The convergence was determined by the stopping criteria $\|f(w_l)\| \leq 10^{-2}$. The exact location of roots is shown by two white dots within the basins. The dark blue color (left side of basins) represent the convergence region corresponding to the root $w = -1$ and the light blue color (right side of basins) represent the convergence region corresponding to the root $w = 1$.

Figure 1 illustrates the basins for NM (5), NSM (11), HM (12), $PA_{r,s}$ (10), and the proposed method PM (18). It is evident from Figure 1b,d that methods NSM and $PA_{3,1}$ exhibit local convergence, while the other methods exhibit the global convergence. Moreover, Figure 1e,f illustarte broader and lighter basins of attraction because of higher convergence order of $PA_{2,2}$ and PM as compared to the basins of NM and HM as shown in Figure 1a,c. One can notice the symmetry in convergence region of each method. The convergence regions are symmetric about imaginary axis.

After achieving global convergence of the proposed method, it is important to examine certain critical factors related to matrix iterations, including convergence rate and numerical stability.
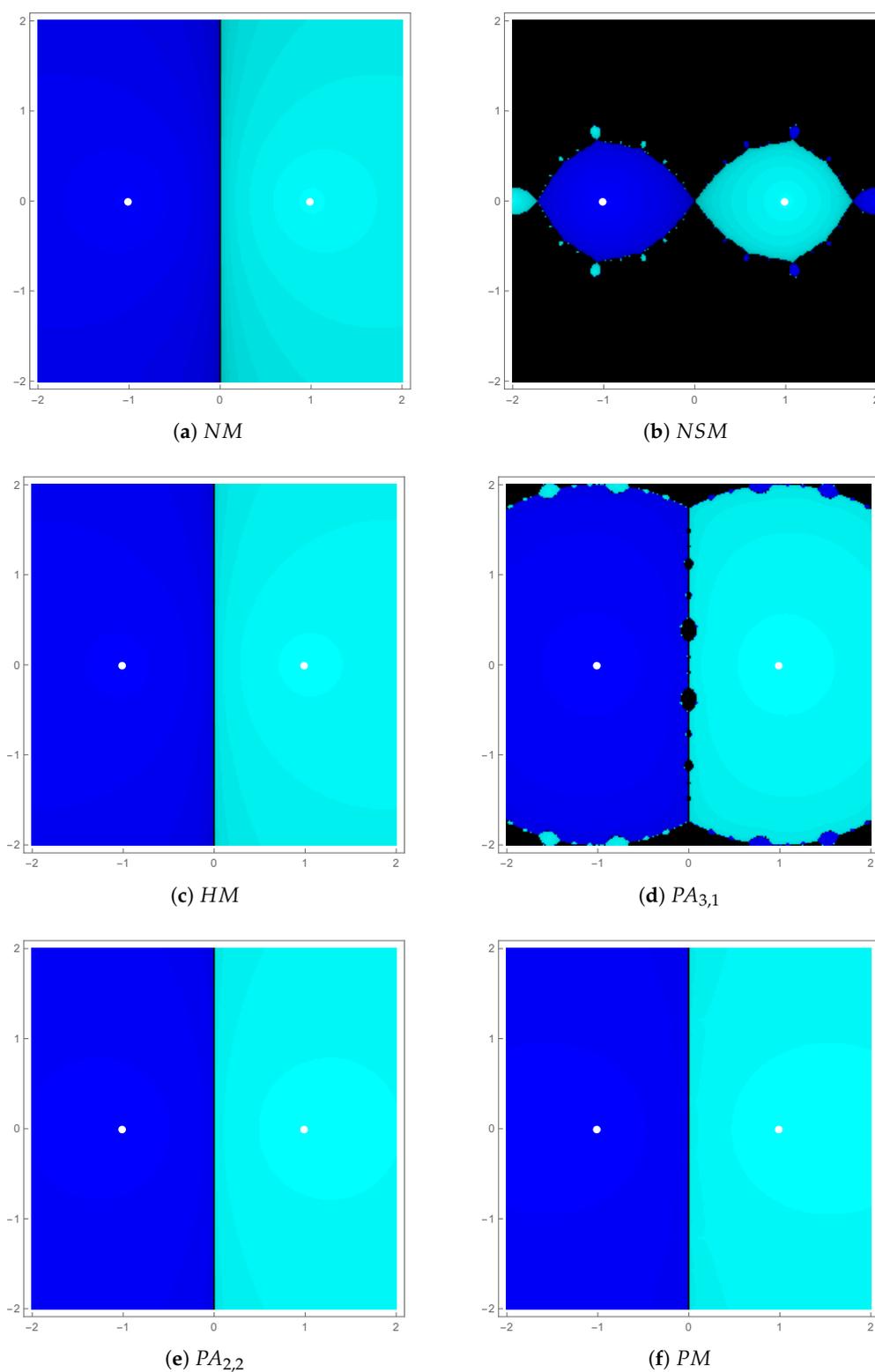
Figure 1. Attraction basins of various methods for equation $w^2 - 1 = 0$.

3.1. Convergence Analysis

Theorem 1. *Suppose that the initial matrix $W_0 = A \in \mathbb{C}^{n \times n}$ has no eigenvalue on the imaginary axis, then the sequence of matrices $\{W_l\}_{l=0}^{\infty}$ generated from (18) converges to $\text{sign}(A) = S$.*

Proof. Let $W = PJP^{-1}$ be Jordan decomposition of matrix $W \in \mathbb{C}^{n \times n}$ and consider

$$R(W) = PR(J)P^{-1}. \tag{21}$$

where R is the rational function linked to (18). Consequently, this will lead to the mapping of an eigenvalue λ of W_l to an eigenvalue $R(\lambda)$ of W_{l+1}. Also, R will satisfy the properties given as follows:

1. $sign(R(w)) = sign(w) \ \forall \ w \in \mathbb{C}$.
2. $w_{l+1} = R(w_l)$ with $w_0 = w$ converges to $sign(w)$, where $Re(w) \neq 0$.

To achieve this objective, consider $A = P\Lambda P^{-1}$ as the Jordan decomposition of A that can be modified as [2]

$$P^{-1}AP = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{22}$$

where P is a non singular matrix while Jordan blocks C and N are associated to the eigenvalues in $\mathbb{C}^-$ and $\mathbb{C}^+$ respectively. Let the principal diagonals of blocks C and N have values $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $\lambda_{p+1}, \lambda_{p+2}, \ldots, \lambda_n$ respectively. Then

$$sign(A) = P \begin{bmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{bmatrix} P^{-1}. \tag{23}$$

Therefore,

$$\begin{aligned} sign(\Lambda) = sign(P^{-1}AP) &= P^{-1}sign(A)P \\ &= \begin{bmatrix} sign(\lambda_1) & & & & & \\ & \ddots & & & & \\ & & sign(\lambda_p) & & & \\ & & & sign(\lambda_{p+1}) & & \\ & & & & \ddots & \\ & & & & & sign(\lambda_n) \end{bmatrix} \end{aligned} \tag{24}$$

Now, define a sequence $D_l = P^{-1}W_l P$ with $D_0 = P^{-1}AP$. Then from (18), we get

$$D_{l+1} = [D_l(21I + 50D_l^2 + 9D_l^4)][4I + 45D_l^2 + 30D_l^4 + D_l^6]^{-1}. \tag{25}$$

Mathematical induction states that all D_l's for $l = 1, 2, 3\ldots$ will be diagonal if D_0 is diagonal. When D_0 is not diagonal, the proof may be considered in the similar methodology as stated at the end of the proof.

Now, (25) can be written for scalar case $f(w) = w^2 - 1 = 0$ as follows:

$$d_{l+1}^i = \frac{d_l^i(21 + 50d_l^{i2} + 9d_l^{i4})}{4 + 45d_l^{i2} + 30d_l^{i4} + d_l^{i6}}, \tag{26}$$

where $d_l^i = (D_l)_{i,i}$ and $1 \leq i \leq n$. Furthermore, $sign(\lambda_i) = s_i$, with $s_i = \pm 1$ and therefore from (26),

$$\frac{d_{l+1}^i - s_i}{d_{l+1}^i + s_i} = \left(\frac{-s_i + d_l^i}{s_i + d_l^i} \right)^5 \left(\frac{4s_i - d_l^i}{4s_i + d_l^i} \right). \tag{27}$$

The factor $\frac{4s_i - d_l^i}{4s_i + d_l^i}$ is bounded for $1 \leq i \leq n$ and does not affect the convergence. Furthermore, by selecting the right initial matrix $W_0 = A$ and $\left| \frac{d_0^i - s_i}{d_0^i + s_i} \right| < 1$, we get

$$\lim_{l \to \infty} \left| \frac{d_{l+1}^i - s_i}{d_{l+1}^i + s_i} \right| = 0, \tag{28}$$

implying that $\lim_{l \to \infty}(d_l^i) = s_i = sign(\lambda_i)$ and hence $\lim_{l \to \infty}(D_l) = sign(\Lambda)$.

In the case when D_0 is not a diagonal matrix, it is essential that we investigate the relation between the eigenvalues of W_i's which was briefly explained at the starting of the proof.

$$\lambda_{l+1}^i = [\lambda_l^i(21I + 50\lambda_l^{i^2} + 9\lambda_l^{i^4})][4I + 45\lambda_l^{i^2} + 30\lambda_l^{i^4} + \lambda_l^{i^6}]^{-1}. \tag{29}$$

In a comparable approach, it is clear from (29) that λ_i's will converge to $s_i = \pm 1$, i.e.,

$$\lim_{l \to \infty} \left| \frac{\lambda_{l+1}^i - s_i}{\lambda_{l+1}^i + s_i} \right| = 0. \tag{30}$$

As a result, it is easy to conclude that

$$\lim_{l \to \infty} W_l = P(\lim_{l \to \infty} D_l)P^{-1} = Psign(\Lambda)P^{-1} = sign(A), \tag{31}$$

which finishes the proof. $\square$

Theorem 2. *Considering the assumptions of Theorem 1, the matrix sequence $\{W_l\}_{l=0}^{\infty}$ generated from (18) has convergence rate five to $sign(A) = S$.*

Proof. Since, W_l depends on matrix $A \; \forall \; l \geq 0$ and the matrix A commutes with S, so W_l also commutes with $S \; \forall \; l \geq 0$. Taking into consideration the substitution

$$B_l = 4I + 45W_l^2 + 30W_l^4 + W_l^6, \tag{32}$$

we get,

$$
\begin{aligned}
W_{l+1} - S &= [W_l(21I + 50W_l^2 + 9W_l^4)]B_l^{-1} - S \\
&= [21W_l + 50W_l^3 + 9W_l^5 - SB_l]B_l^{-1} \\
&= [21W_l + 50W_l^3 + 9W_l^5 - 4S - 45SW_l^2 - 30SW_l^4 - SW_l^6]B_l^{-1} \\
&= [-4S^5 + 21S^4W_l - 45S^3W_l^2 + 50S^2W_l^3 - 30SW_l^4 + 9W_l^5 - SW_l^6]B_l^{-1} \\
&= (W_l - S)^5(4I - SW_l)B_l^{-1}
\end{aligned}
\tag{33}
$$

Apply any type of matrix norm to the both sides of (33),

$$\|W_{l+1} - S\| \leq \left(\|4I - SW_l\| \left\| B_l^{-1} \right\| \right) \|W_l - S\|^5. \tag{34}$$

This completes the proof. $\square$

3.2. Stability Issues

Theorem 3. *Considering the assumptions of Theorem 1, the sequence of matrix iterations $\{W_l\}_{l=0}^{\infty}$ generated from (18) is stable asymptotically.*

Proof. The iterate obtained from (18) is a function of A because of $W_0 = A$, and hence commutes with A. Let Δ_l be the perturbation generated at l^{th} iterate in (18). Then

$$\widetilde{W}_l = W_l + \Delta_l. \tag{35}$$

Based on the results of the error analysis of order one, we may consider that $(\Delta_l)^j \approx 0$, $j \geq 2$. As long as Δ_l is small enough, this usual adjustment will work. Also suppose that $W_l \approx sign(A) = S$ for larger l, then we get

$$\begin{aligned}
\widetilde{W}_{l+1} &= [\widetilde{W}_l(21I + 50\widetilde{W}_l^2 + 9\widetilde{W}_l^4)][4I + 45\widetilde{W}_l^2 + 30\widetilde{W}_l^4 + \widetilde{W}_l^6]^{-1} \\
&= [(W_l + \Delta_l)(21I + 50(W_l + \Delta_l)^2 + 9(W_l + \Delta_l)^4)][4I + 45(W_l + \Delta_l)^2 \\
&\quad + 30(W_l + \Delta_l)^4 + (W_l + \Delta_l)^6]^{-1} \\
&\approx [80S + 148\Delta_l + 68S\Delta_l S][80I + 108S\Delta_l + 108\Delta_l S]^{-1} \\
&\approx \left[S + \frac{37}{20}\Delta_l + \frac{17}{20}S\Delta_l S\right]\left[I - \frac{27}{20}S\Delta_l - \frac{27}{20}\Delta_l S\right] \\
&\approx S + \frac{1}{2}\Delta_l - \frac{1}{2}S\Delta_l S.
\end{aligned} \tag{36}$$

Here, the identity [18] given by

$$(B + C)^{-1} = B^{-1} - B^{-1}CB^{-1} + O(C^2). \tag{37}$$

has been applied, where B is any nonsingular matrix and C is any matrix. Also, considering $\Delta_{l+1} = \widetilde{W}_{l+1} - W_{l+1}$, we get

$$\Delta_{l+1} \approx \frac{1}{2}\Delta_l - \frac{1}{2}S\Delta_l S. \tag{38}$$

At this point, it is possible to deduce that Δ_{l+1} is bounded, i.e.,

$$\|\Delta_{l+1}\| \leq \frac{1}{2}\|\Delta_0 - S\Delta_0 S\|, \tag{39}$$

which completes the proof. $\square$

4. Extension to Determine Generalized Eigenvalues of a Matrix Pencil

Let $A, B \in \mathbb{C}^{n \times n}$. Then $\lambda \in \mathbb{C}$ is said to be a generalized eigenvalue of a matrix pencil $A - \lambda B$ if there exists a non-zero vector $\mathbf{x} \in \mathbb{C}^n$ such that

$$A\mathbf{x} = \lambda B\mathbf{x}. \tag{40}$$

Here $\mathbf{x}$ will be a generalized eigenvector corresponding to λ and (40) is referred as a generalized eigenvalue problem. This can be determined by finding the zeros of characteristic polynomial $det(A - \lambda B)$.

Complex problem solving in physics was the first scenario for the appearance of the characteristic polynomial and the eigenproblem. For a recent coverage of the topic one can see [19]. When more generality is needed, generalized eigenvalue problem is taking the place of the classical approach [20].

Problem (40) arises in several fields of numerical linear algebra. However, the dimensions of matrices A and B are typically quite enormous, hence further complicating the situation. There are a lot of different approaches that claim to be able to solve (40) including the necessary eigenvectors in a smaller subspace or series of matrix pencils [21–23] or using matrix decomposition [24–26]. Our interest is in employing the matrix sign function which involves evaluating certain matrix sign computations based on the given matrix pencil to compute the required decomposition and hence to solve the problem (40).

To solve the problem (40) for the case of regular matrix pencils using matrix sign function, we consider rank revealing QR ($RRQR$) decomposition [27] after computing sign of two matrices depending on the given matrix pencil.

Theorem 4. *Let $A, B \in \mathbb{C}^{n \times n}$ such that $A - \lambda B$ be a regular matrix pencil and let U_1 and U_2 are defined as*

$$U_1 = \frac{I - S_1}{2}, \ U_2 = \frac{I - S_2}{2}, \tag{41}$$

where S_1 and S_2 are sign functions given by

$$
\begin{aligned}
S_1 &= sign((A - B)^{-1}(A + B)), \\
S_2 &= sign((A + B)(A - B)^{-1}).
\end{aligned}
\tag{42}
$$

If $U_1 \Pi_1 = Q_1 R_1$ and $U_2 \Pi_2 = Q_2 R_2$ are RRQR decompositions of U_1 and U_2 such that $Q_2^* A Q_1$ and $Q_2^* B Q_1$ are upper triangular, then

$$\lambda_i = \frac{\alpha_{ii}}{\beta_{ii}}, \tag{43}$$

where λ_i's are eigenvalues of $A - \lambda B$, α_{ii} and β_{ii} are the diagonal entries of $Q_2^* A Q_1$ and $Q_2^* B Q_1$ respectively for $i = 1, 2, \ldots, n$. Moreover, if B is non-invertible, then $\beta_{ii} = 0$ for some i which implies that some eigenvalues of $A - \lambda B$ will be infinity.

Let τ denotes tolerance value and *max* be the maximum iterations. Then, in accordance with the proposed method (18) and Theorem 4, we propose Algorithm 1 to find out the eigenvalues of a regular matrix pencil $A - \lambda B$.

Algorithm 1 Evaluating generalized eigenvalues of $A - \lambda B$

Require: A, B, max, τ
Ensure: $S_1, S_2, \lambda_i's$
1: Initialize $W_0 = (A - B)^{-1}(A + B)$.
2: **while** $l \leq max$ and $\|W_l^2 - I\|_* > \tau$ **do**
3:      Apply method (18).
4:      **return** $W_\infty = sign((A - B)^{-1}(A + B)) = S_1$.
5: **end while**
6: Construct *RRQR* decomposition of $U_1 = \frac{I - S_1}{2}$ as follows:
     $U_1 \Pi_1 = Q_1 R_1$.
7: Initialize $W_0 = (A + B)(A - B)^{-1}$.
8: Repeat step 2–5 to **return** $W_\infty = sign((A + B)(A - B)^{-1}) = S_2$.
9: Construct *RRQR* decomposition of $U_2 = \frac{I - S_2}{2}$ as follows:
     $U_2 \Pi_2 = Q_2 R_2$.
10: Compute two matrices $Q_2^* A Q_1 = [\alpha_{ij}]$ and $Q_2^* B Q_1 = [\beta_{ij}]$, where $1 \leq i, j \leq n$.
11: **if** $\alpha_{ij} = \beta_{ij} = 0 \ \forall \ i > j$ **then**
12:      **for** i=1:n **do**
13:          $\lambda_i = \frac{\alpha_{ii}}{\beta_{ii}}$.
14:      **end for**
15: **end if**

---

5. Computational Complexity

The computational efficiency index (CEI) [28] of an iterative method can be defined as follows:

$$CEI = p^{\frac{1}{C}}. \tag{44}$$

Here p denotes convergence order and C denotes computational cost per iteration.

The approximate cost of matrix multiplications in current programming packages is 2.373 [29]. For a matrix of size n, we therefore consider that the cost of one matrix-matrix multiplication is $n^{\frac{2373}{1000}}$ and cost of one matrix inverse is $\frac{2n^3}{3}$. Additionally, we do not take into account the cost per iteration of norm computation (in ANM (6)), addition

and subtraction. Then, *CEI*'s for the methods *NM* (5), *ANM* (6), *HM* (12) and *PM* (18) to compute $sign(A)$ are given in Table 1.

Table 1. *CEI*'s for various methods along with their convergence orders.

Method	Convergence Order	CEI
NM	2	$2^{\frac{1}{n^{2373/1000} + \frac{2n^3}{3}}}$
ANM	$\sqrt{3}+1$	$(\sqrt{3}+1)^{\frac{1}{n^{2373/1000} + \frac{2n^3}{3}}}$
HM	3	$3^{\frac{1}{3n^{2373/1000} + \frac{2n^3}{3}}}$
PM	5	$5^{\frac{1}{5n^{2373/1000} + \frac{2n^3}{3}}}$

Figure 2 presents a comparison of the efficiencies of several approaches, demonstrating the greater efficiency index of the suggested method.
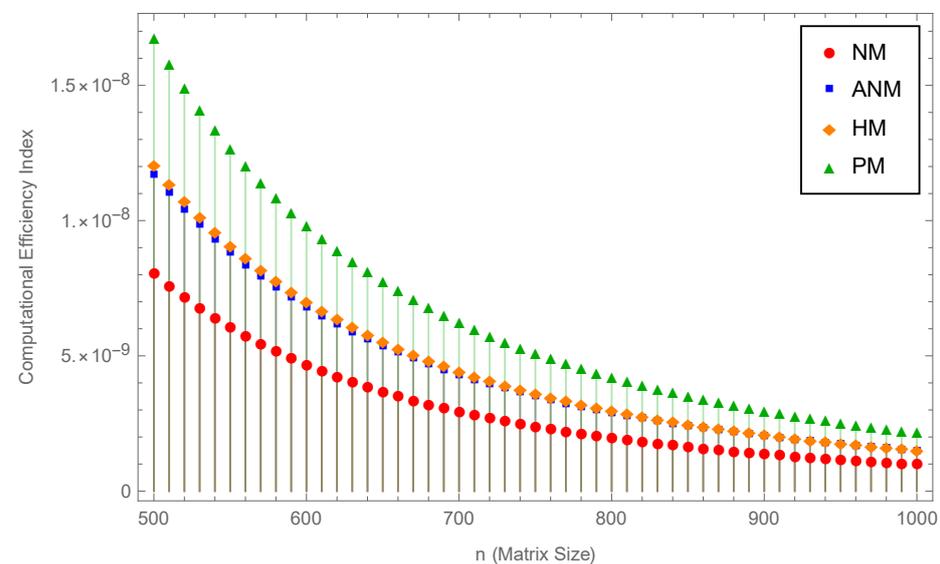

Figure 2. Comparative analysis of efficiency indices for different methods.

6. Numerical Aspects

In this section, we have examined a variety of examples to evaluate our proposed method by comparing it with various existing techniques. For this study, we have assumed the following stopping criteria:

$$\|W_l^2 - I\|_* \leq \tau, \tag{45}$$

where τ stands for the tolerance value and $\|.\|_*$ will be suitable matrix norm. Also, l_∞ and l_2 norms are considered for real and complex input matrices, respectively [14].

For the sake of comparison, we only consider the methods that exhibit global convergence and we do not include any methods that have local convergence. The methods that are being compared will be *NM* (5), *ANM* (6), *HM* (12), $PA_{2,2}$ ((10) for $r = s = 2$), *LM* (13), *PM* (18) and *APM* (20) with spectral scaling (19). All simulations are done in Mathematica using system specifications "Windows 11 Home Intel(R) Core(TM) i7-1255U CPU running at 1.70 GHz processor having 16.0 RAM (64-bit operating system)".

Example 1. *Here, we perform the eigenvalue clustering of a random real* 500×500 *matrix given as*

```
SeedRandom[100]; A = RandomReal[{-5,5},{500,500}];
```

The outcomes are shown in Figure 3 using the stopping criteria (45) with l_∞ norm and $\tau = 10^{-4}$. We can see that eigenvalues of W_0 are scattered at the initial stage as shown

in Figure [3]a, while proceeding for the next iterations W_i in Figure [3]b–g for $i = 1, 2, \ldots, 6$ respectively using PM (18), the eigenvalues are converging towards 1 or -1. Then in Figure [3]h, all eigenvalues converge to 1 or -1 (based on the stopping criteria) implying that the sequence of matrices $\{W_l\}_{l=0}^{\infty}$ converges to $sign(A)$. The result of Theorem [1] is therefore confirmed.
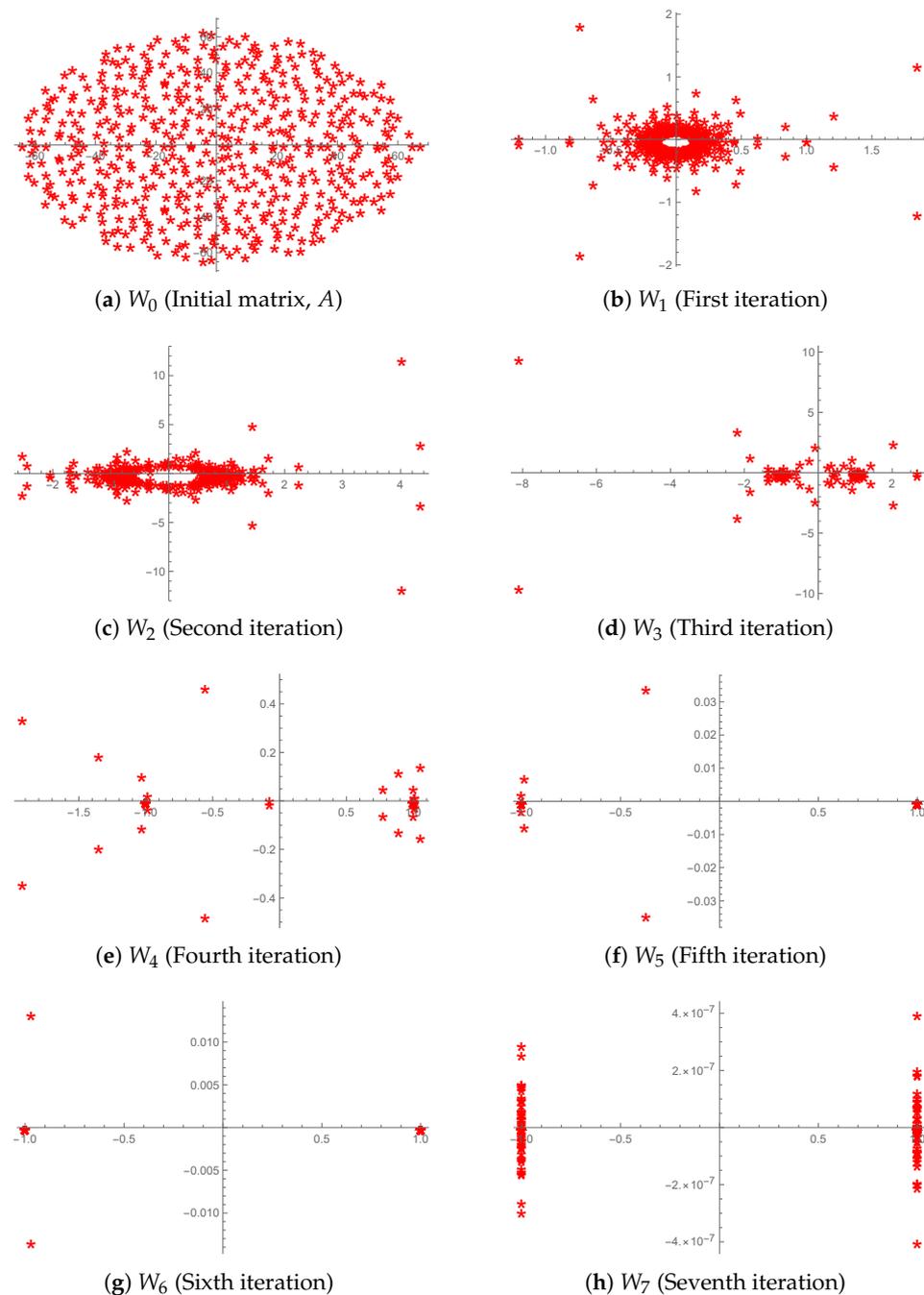

(**a**) W_0 (Initial matrix, A)

(**b**) W_1 (First iteration)

(**c**) W_2 (Second iteration)

(**d**) W_3 (Third iteration)

(**e**) W_4 (Fourth iteration)

(**f**) W_5 (Fifth iteration)

(**g**) W_6 (Sixth iteration)

(**h**) W_7 (Seventh iteration)

Figure 3. Eigenvalue clustering of 500×500 matrix at every iterate in Example [1].

Example 2. *In this example, convergence of many different methods for computing $sign(A)$ has been compared, where A is a random 1000×1000 complex square matrix given as follows:*

```
SeedRandom[101];
A = RandomComplex[{-1 - 1.5 I, 1 + 1.5 I}, {1000, 1000}]
```

The comparison is shown in Figure 4 with $\tau = 10^{-4}$ and l_2 norm which illustrates the results with regard to the number of iterations and the absolute error and demonstrates the consistent behavior of both *PM* and *APM*.
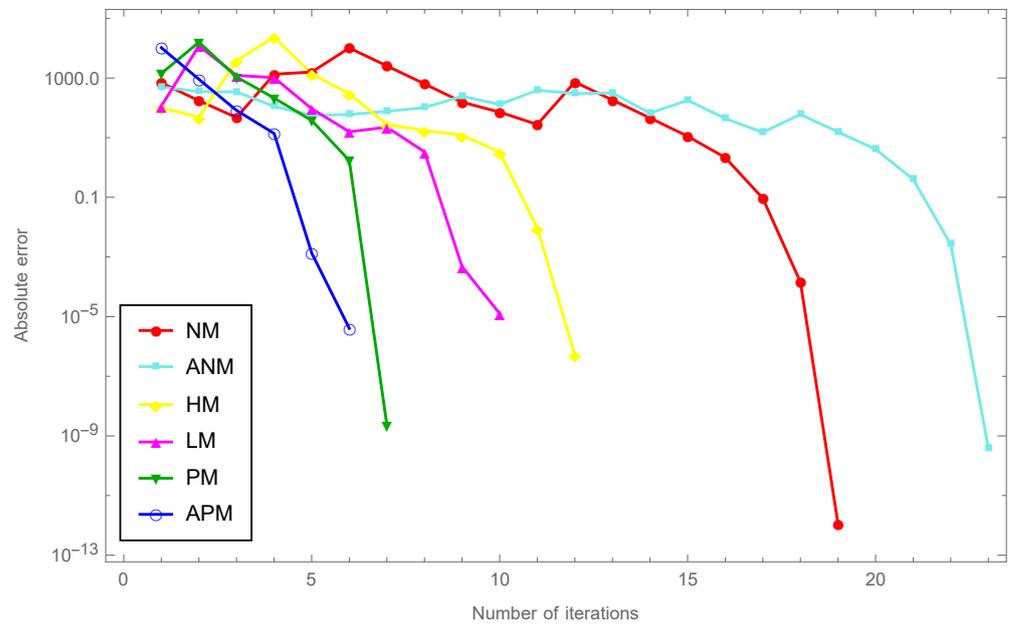

Figure 4. Convergence history of different methods for Example 2.

Example 3. *This example presents a comparison on convergence of many different methods for computing* $sign(A)$*, where* A *is a random* 2000×2000 *real square matrix given as*

```
SeedRandom[111];
A = RandomReal[{-25, 25}, {2000, 2000}];
```

The comparison is presented in Figure 5 with $\tau = 10^{-4}$ and l_∞ norm which illustrates the results with respect to the number of iterations and the absolute error and demonstrates the consistent behavior of both *PM* and *APM* for evaluating $sign(A)$.
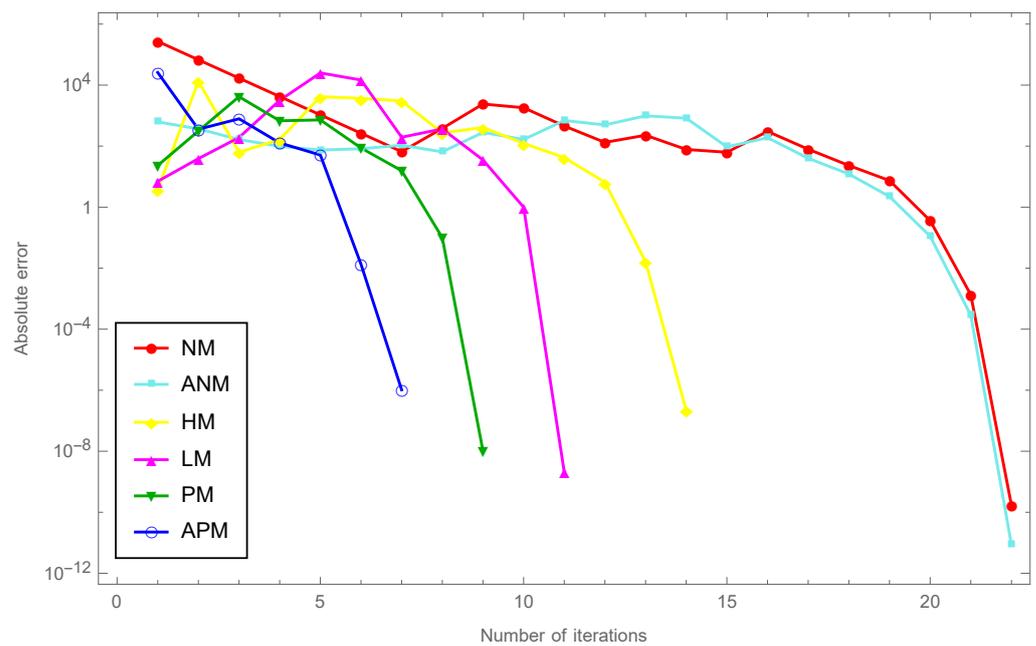

Figure 5. Convergence history of different methods for Example 3.

Example 4. *Here, we examine the convergence of various methods with regard to the iteration number and the timings of CPU for evaluating the sign function of ten random matrices:*

```
SeedRandom[121]; number = 10; tolerance = 10^(-5);
Table[A[j] = RandomComplex[{-3 - 2 I, 3 + 2 I}, {50j, 50j}];, {j, number}];
```

The comparative analysis based on iteration number and CPU timings for matrices of various sizes, are presented in Table 2 and Table 3 respectively using $\tau = 10^{-5}$ and l_2-norm. The implementation of new methods (PM) and (APM) result in an improvement in the average iteration number and the amount of time spent by CPU. The outcomes demonstrate that the suggested method (PM) is significantly better as compared to other fifth order methods like $PA_{2,2}$ and LM.

Table 2. Comparative analysis for the iteration number in Example 4.

Matrix	NM	ANM	HM	LM	$PA_{2,2}$	PM	APM
$A_{50 \times 50}$	12	11	8	6	6	5	4
$A_{100 \times 100}$	15	13	9	7	7	6	5
$A_{150 \times 150}$	15	15	10	8	7	6	5
$A_{200 \times 200}$	15	15	9	8	7	6	5
$A_{250 \times 250}$	15	17	10	8	7	6	5
$A_{300 \times 300}$	14	15	9	7	6	6	5
$A_{350 \times 350}$	15	16	10	8	7	6	5
$A_{400 \times 400}$	17	19	11	10	7	7	5
$A_{450 \times 450}$	14	15	9	7	6	6	5
$A_{500 \times 500}$	18	21	12	9	8	7	6
Mean	15	15.7	9.7	7.8	6.8	6.1	5

Table 3. Comparative analysis for the CPU-time (in seconds) in Example 4.

Matrix	NM	ANM	HM	LM	$PA_{2,2}$	PM	APM
$A_{50 \times 50}$	0.03	0.02	0.01	0.01	0.02	0.01	0.01
$A_{100 \times 100}$	0.07	0.15	0.05	0.05	0.09	0.05	0.10
$A_{150 \times 150}$	0.16	0.23	0.12	0.15	0.21	0.10	0.24
$A_{200 \times 200}$	0.29	0.40	0.19	0.26	0.34	0.18	0.40
$A_{250 \times 250}$	0.47	0.67	0.32	0.39	0.53	0.25	0.60
$A_{300 \times 300}$	0.58	0.87	0.44	0.55	0.63	0.44	1.02
$A_{350 \times 350}$	0.82	1.29	0.70	0.88	1.21	0.63	1.44
$A_{400 \times 400}$	1.20	2.04	1.46	1.78	1.62	0.99	1.70
$A_{450 \times 450}$	1.31	2.06	1.08	1.40	1.87	1.16	2.29
$A_{500 \times 500}$	2.06	3.58	1.91	2.32	3.07	1.60	3.32
Mean	0.70	1.13	0.63	0.78	0.96	0.54	1.11

Now, in Examples 5 and 6, we aim to evaluate generalized eigenvalues of matrix pencils using matrix sign function for which we use Algorithm 1 for different methods by replacing (18) with existing methods in Step 3.

Example 5. *Let $A - \lambda B$ be a matrix pencil, where matrices A and B are given by [30]*

$$
A = \begin{bmatrix}
\frac{99}{100} & \frac{1}{100} & 0 & \cdots & 0 \\
0 & \frac{98}{100} & \ddots & & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \frac{1}{100} & \frac{1}{100} \\
0 & \cdots & 0 & 0 & \frac{0}{100}
\end{bmatrix}_{100 \times 100}, \quad
B = \begin{bmatrix}
0 & & & \\
& \ddots & & \\
& & 0 & \\
& & & I_{20}
\end{bmatrix}_{100 \times 100},
$$

where I_{20} is the identity matrix of size 20.

The eigenvalues of given pencil $A - \lambda B$ are $\lambda_i = \frac{i-1}{100}$ for $i = 1, 2, \ldots 20$ and remaining eigenvalues are infinite. We verify these eigenvalues using proposed method and Algorithm 1. We compute the matrices $Q_2^* A Q_1 = [\alpha_{ij}]_{100 \times 100}$ and $Q_2^* B Q_1 = [\beta_{ij}]_{100 \times 100}$ and then calculate λ_i's from Algorithm 1. The outcomes are presented in Table 4 for $i = 1, 2, \ldots 20$.

Table 4. Eigenvalues (λ_i's) of matrix pencil $A - \lambda B$ in Example 5.

i	1	2	3	4	5	6	7	8	9	10
α_{ii}	0.189762	0.18	-0.17	0.16	0.15	-0.14	0.13	-0.12	-0.11	0.1
β_{ii}	0.99875	1	-1	1	1	-1	1	-1	-1	1
λ_i	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.12	0.11	0.1

i	11	12	13	14	15	16	17	18	19	20
α_{ii}	-0.09	0.08	-0.07	0.05	-0.03	0.02	0.01	0	-0.06	0.04
β_{ii}	-1	1	-1	1	-1	1	1	-1	-1	1
λ_i	0.09	0.08	0.07	0.05	0.03	0.02	0.01	0	0.06	0.04

Moreover, $\beta_{ii} = 0$ for $21 \le i \le 100$ which implies that $\lambda_i = \infty$ for $21 \le i \le 100$. Hence this example verifies the results given in Theorem 4.

Example 6 (Bounded finline dielectric waveguide problem). *The generalized eigenvalue problem* (40) *arises in the finite element analysis for finding the propagating modes of a rectangular waveguide. We consider three cases for the pairs of matrices A and B as BFW62A, BFW62B; BFW398A, BFW398B and BFW782A, BFW782B from the Matrix Market [31].*

Table 5 shows the outcomes depending on the iteration numbers l_1 and l_2 for computing $S_1 = sign((A - B)^{-1}(A + B))$ and $S_2 = sign((A + B)(A - B)^{-1})$ with corresponding absolute errors $\epsilon_{l_1} = \|W_{l_1}^2 - I\|_\infty$ and $\epsilon_{l_2} = \|W_{l_2}^2 - I\|_\infty$ using stopping criterion (45) with $\tau = 10^{-10}$. Also, total computational time in seconds taken for computing S_1 and S_2 has been compared in Figure 6. The outcomes demonstrate that the suggested method is better with respect to errors and timings.

Table 5. Comparison of results obtained from various methods in Example 6.

| Matrices	Method	l_1	ϵ_{l_1}	l_2	ϵ_{l_2}
BFW62A, BFW62B	NM	2	5.79607×10^{-11}	2	5.79607×10^{-11}
	HM	2	4.44089×10^{-16}	2	4.44089×10^{-16}
	$PA_{2,2}$	1	1.65201×10^{-13}	1	1.64757×10^{-13}
	LM	1	9.80327×10^{-13}	1	9.81437×10^{-13}
	PM	1	9.90319×10^{-14}	1	1.00364×10^{-13}
BFW398A, BFW398B	NM	2	6.18883×10^{-12}	2	6.18883×10^{-12}
	HM	2	4.44089×10^{-16}	2	4.44089×10^{-16}
	$PA_{2,2}$	1	1.50990×10^{-14}	1	1.50990×10^{-14}
	LM	1	9.08162×10^{-14}	1	9.23706×10^{-14}
	PM	1	1.11022×10^{-14}	1	1.11022×10^{-14}
BFW782A, BFW782B	NM	2	5.82201×10^{-13}	2	5.82201×10^{-13}
	HM	2	4.44089×10^{-16}	2	4.44089×10^{-16}
	$PA_{2,2}$	1	6.43929×10^{-15}	1	3.99680×10^{-15}
	LM	1	7.99361×10^{-15}	1	7.99361×10^{-15}
	PM	1	5.32907×10^{-16}	1	5.32907×10^{-16}
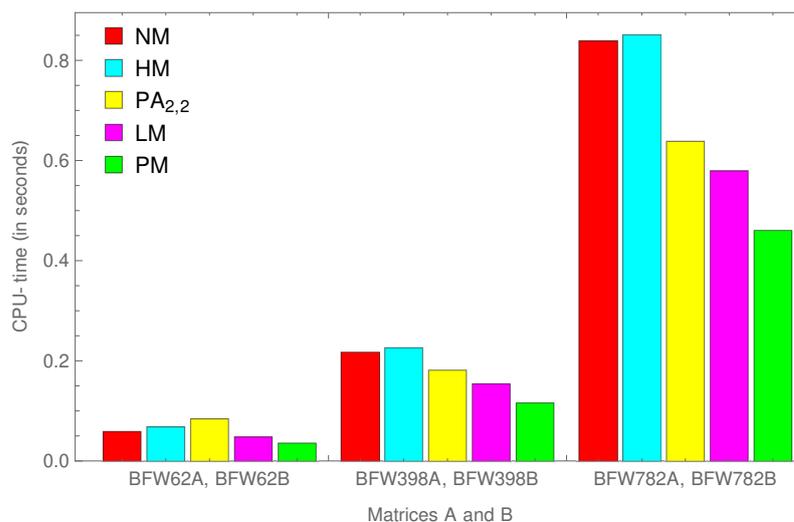
Figure 6. Comparison of CPU time for various methods to compute S_1 and S_2 in Example 6.

7. Conclusions

This paper has discussed an iterative approach for evaluation of sign of a matrix. Our proposed iterative method attains a convergence order of five, showcasing its efficiency in evaluation of the sign of complex matrices. Through comprehensive convergence analysis discussed globally and the demonstration of asymptotic stability, we have established the robustness and reliability of the proposed approach. Then an algorithm has been designed for determining the generalized eigenvalues of a regular matrix pencil utilizing matrix sign computation. Furthermore, extensive numerical experiments performed on matrices of different sizes have validated the effectiveness and superiority of our method over existing techniques.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; software, P.S.; validation, M.K. and V.S.; formal analysis, L.J.; investigation, P.S.; resources, L.J.; data curation, L.J.; writing—original draft preparation, P.S.; writing—review and editing, L.J.; visualization, V.S.; supervision, M.K. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Roberts, J.D. Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. *Int. J. Control* **1980**, *32*, 677–687. [CrossRef]
2. Higham, N.J. *Functions of Matrices: Theory and Computation*; SIAM: Philadelphia, PA, USA, 2008. [CrossRef]
3. Kenney, C.; Laub, A.J. Rational iterative methods for the matrix sign function. *SIAM J. Matrix Anal. Appl.* **1991**, *12*, 273–291. [CrossRef]
4. Kenney, C.S.; Laub, A.J. The matrix sign function. *IEEE Trans. Autom. Control* **1995**, *40*, 1330–1348. [CrossRef]
5. Gomilko, O.; Greco, F.; Ziętak, K. A Padé family of iterations for the matrix sign function and related problems. *Numer. Linear Algebra Appl.* **2012**, *19*, 585–605. [CrossRef]
6. Schulz, G. Iterative berechung der reziproken matrix. *J. Appl. Math. Mech. Angew. Math. Mech.* **1933**, *13*, 57–59. [CrossRef]
7. Momenzadeh, M.; Lotfi, T. New iterative methods for finding matrix sign function: Derivation and application. *Comput. Appl. Math.* **2019**, *38*, 53. [CrossRef]
8. Ullah, M.Z.; Alaslani, S.M.; Mallawi, F.O.; Ahmad, F.; Shateyi, S.; Asma, M. A fast and efficient Newton-type iterative scheme to find the sign of a matrix. *AIMS Math.* **2023**, *8*, 19264–19274. [CrossRef]

9. Wang, X.; Cao, Y. A numerically stable high-order Chebyshev-Halley type multipoint iterative method for calculating matrix sign function. *AIMS Math.* **2023**, *8*, 12456–12471. [CrossRef]
10. Liu, T.; Ullah, M.Z.; Alshahrani, K.M.A.; Shateyi, S. From fractal behavior of iteration methods to an efficient solver for the sign of a matrix. *Fractal Fract.* **2023**, *7*, 32. [CrossRef]
11. Benner, P.; Quintana-Ortí, E.S. Solving stable generalized Lyapunov equations with the matrix sign function. *Numer. Algorithms* **1999**, *20*, 75–100. [CrossRef]
12. Howland, J.L. The sign matrix and the separation of matrix eigenvalues. *Linear Algebra Its Appl.* **1983**, *49*, 221–232. [CrossRef]
13. Soleymani, F.; Kumar, A. A fourth-order method for computing the sign function of a matrix with application in the Yang–Baxter-like matrix equation. *Comput. Appl. Math.* **2019**, *38*, 64. [CrossRef]
14. Soheili, A.R.; Toutounian, F.; Soleymani, F. A fast convergent numerical method for matrix sign function with application in SDEs. *J. Comput. Appl. Math.* **2015**, *282*, 167–178. [CrossRef]
15. Kenney, C.; Laub, A.J. On scaling Newton's method for polar decomposition and the matrix sign function. *SIAM J. Matrix Anal. Appl.* **1992**, *13*, 688–706. [CrossRef]
16. Soleymani, F.; Stanimirović, P.S.; Stojanović, I. A novel iterative method for polar decomposition and matrix sign function. *Discrete Dyn. Nat. Soc.* **2015**, *2015*, 649423. [CrossRef]
17. Iannazzo, B. Numerical Solution of Certain Nonlinear Matrix Equations. Ph.D. Thesis, Dipartimento di Matematica, Università di Pisa, Pisa, Italy, 2007.
18. Henderson, H.V.; Searle, S.R. On deriving the inverse of a sum of matrices. *SIAM Rev.* **1981**, *23*, 53–60. [CrossRef]
19. Jäntschi, L. Eigenproblem Basics and Algorithms. *Symmetry* **2023**, *15*, 2046. [CrossRef]
20. Fischer, M.; Kostrzewa, B.; Ostmeyer, J.; Ottnad, K.; Ueding, M.; Urbach, C. On the generalised eigenvalue method and its relation to Prony and generalised pencil of function methods. *Eur. Phys. J. A* **2020**, *56*, 206. [CrossRef]
21. Ruhe, A. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra Appl.* **1984**, *58*, 391–405. [CrossRef]
22. Gallivan, K.; Grimme, E.; Van Dooren, P. Asymptotic waveform evaluation via a Lanczos method. *Appl. Math. Lett.* **1994**, *7*, 75–80. [CrossRef]
23. Gallivan, K.; Grimme, E.; Van Dooren, P. A rational Lanczos algorithm for model reduction. *Numer. Algor.* **1996**, *12*, 33–63. [CrossRef]
24. Moler, C.B.; Stewart, G.W. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Num. Anal.* **1973**, *10*, 241–256. [CrossRef]
25. Bai, Z.; Demmel, J. Using the matrix sign function to compute invariant subspaces. *SIAM J. Matrix Anal. Appl.* **1998**, *19*, 205–225. [CrossRef]
26. Sun, X.; Quintana-Ortí, E. Spectral division methods for block generalized Schur decompositions. *Math. Comput.* **2004**, *73*, 1827–1847. [CrossRef]
27. Quintana-Ortí, G.; Sun, X.; Bischof, C.H. A BLAS-3 Version of the QR Factorization with Column Pivoting. *SIAM J. Sci. Comp.* **1998**, *19*, 1486–1494. [CrossRef]
28. Soleymani, F.; Khaksar Haghani, F.; Shateyi, S. Several numerical methods for computing unitary polar factor of a matrix. *Adv. Differ. Equ.* **2016**, *2016*, 4. [CrossRef]
29. Le Gall, F. Powers of tensors and fast matrix multiplication. In Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation, Kobe, Japan, 23–25 July 2014; ACM: New York, NY, USA, 2014; pp. 296–303. [CrossRef]
30. Sakurai, T.; Sugiura, H. A projection method for generalized eigenvalue problems using numerical integration. *J. Comput. Appl. Math.* **2003**, *159*, 119–128. [CrossRef]
31. Matrix Market. Available online: https://math.nist.gov/MatrixMarket/ (accessed on 15 January 2024).